Submitted in part fulfilment of the requirements for the degree of Master of Science in Computer Science and Computational Biology

A Hidden Markov Model Framework For Studying Regulation From Chromatin Through RNA

Eran Rosenthal

Supervisor: Dr. Tommy Kaplan

October 2017

Hebrew University of Jerusalem
Faculty of Science - Computer Science and Engineering
I would like to deeply thank Tommy, for being a great supervisor, giving very useful advices and scientific tutoring. Tommy has great ability to express complex concepts in simple, illustrative stories, and he introduced me to different kinds of experimental data and computational approaches. I would like also to thank for the members of the lab, for the interesting ideas and discussions.

I would like to thank for all the people I had worked in collaboration in my master:

- Moshe Oren and Gilad Fuchs for the providing me their experimental data of nascent RNA with RNF20 knockdown for studying the role of monoubiquitylation of H2B.

- Michael Berger and Yuval Malka for the exciting project of studying post-transcriptional 3’UTR cleavage of mRNA transcripts, and for Hanah Margalit and Avital Shimony who helped us in the analysis of miRNA regulation in this project.
Abstract

Information flows from DNA to RNA, through transcription, followed by translation of the RNA transcript to protein. A viable cell must have proper regulation on genes expression. There are varieties of regulation mechanisms along the way, which include regulation on the DNA and expression level of genes, post-transcriptional regulation, and post-translation regulation.

Most of the cells in our body share the same DNA, but they have different functions. Different tissues or cell types have different gene expression patterns. One of the key components in regulation of the gene expression are modifications in the epigenome level. A major component in the epigenome is chromatin - a complex of proteins, mainly histones, which tightly pack the DNA. Chromatin is subject to variety of histone modifications, which correlates with different patterns of gene expression, with each cell type having different characteristic chromatin and transcriptome. Chromatin modifications may affect the DNA accessibility, making regions in the genome accessible to RNA Polymerase (RNA Pol II) for transcribing DNA to RNA. Enhancers, which are enriched by specific chromatin modifications can affect the expression levels of distant genes. Another level of regulation is during the transcription process, for example different genes may vary in their transcription elongation rates, which may affect the dynamics of expression of genes. The transcribed RNA is subject to post-transcriptional processing and regulation, such as 5’-capping, splicing and insertion of poly(A) tail to mRNA. Lastly, post-translational regulation which include protein modifications may affect the proteins behavior and interactions.

These types of regulation (and others) are widely studied. In the last decade the genomics field adopted high-throughput approaches, taking the advantage of the groundbreaking sequencing methods such as RNA-seq and ChIP-seq. This allows us to study system-wide behavior of the cell rather than focusing on specific components.
of the system. Another important aspect of the recent years science is the availability of gigantic amounts of epigenomic data in publicly accessible datasets, and many sophisticated tools for analyzing such data.

There is a growing need for development of computational tools for efficiently analyzing the data, model it and learn patterns in the data to derive interesting insights on the basic principles of biology. As many of the biological signals are processes along time or along space (along the genome or along transcript), it is tempting to use Hidden Markov Models to study them. Here I demonstrate how such framework can enable simple, clear and efficient way for approaching to different kind biological data and studying different questions on the basic principles of biology:

1. Analyzing nascent RNA data, to learn the affect of monoubiquitylation of histone H2B on the transcription elongation rate.

2. Analyzing RNA-seq combined with various treatments (TEX, 3’ pull-down, 5’ pulldown) to establish evidences for stable, autonomous uncapped RNA fragment downstream to the coding region.

3. Analyzing large dataset of Roadmap Epigenomics ChIP-seq for unsupervised annotation of the genome in various cell types.

The models, combined with subsequent analysis, enable us to derive interesting insights and shed light on some of the the regulation processes in the cell.
Contents

Acknowledgments i

Abstract iii

1 Scientific Introduction 1
  1.1 Biological Background 1
  1.1.1 Genome Organization, Chromatin and Histones 1
  1.1.2 Gene Expression and Regulation 2
  1.1.3 Transcription and mRNA Processing 3
  1.1.4 RNA Polyadenylation 3
  1.2 Computational Background 5
  1.2.1 Hidden Markov Model 5
  1.3 Motivation and Objectives 7

2 Histone H2B Monoubiquitylation And Elongation Rates 9
  2.1 Motivation and Objectives 9
  2.1.1 Experimental Schema 10
  2.1.2 Computational Model 10
  2.2 Results 13
  2.2.1 Genome-Wide Annotation of Elongated Regions in 4sU-DRBseq 13
  2.2.2 Estimation of Elongation Rates and Release Time 16
  2.2.3 Segmented Model Estimation of Elongation Rates 19

3 Evidences for Stable, Autonomous Uncapped RNA Fragments 23
  3.1 Motivation and Objectives 23
  3.1.1 Experimental Schema 24
  3.1.2 Computational Model And Analysis 25
3.2 Results .................................................. 27
  3.2.1 Comparing Coding Region to 3’UTR ......................... 27
  3.2.2 Genome Wide Annotation of Cleavage Points ............... 27
  3.2.3 Characterization of Cleavage Point ........................ 32

4 Epigenomic Annotator ............................................. 35
  4.1 Motivation and Objectives ...................................... 35
    4.1.1 Experimental Schema and Data .............................. 36
    4.1.2 Computational Model And Analysis ........................... 37
  4.2 Results .................................................... 38
    4.2.1 Inferred Model: 15 States Model Of The Chromatin ........ 38
    4.2.2 Functional Meaningful Annotation .......................... 42
    4.2.3 Comparing to ChromHMM ................................... 42

5 Discussion ......................................................... 49
  5.1 Monoubiquitylation of Histone H2B ............................. 49
  5.2 Evidences for Stable, Autonomous Uncapped RNA Fragments .. 50
  5.3 Epigenomic Annotator .......................................... 51

Bibliography ......................................................... 53
List of Figures

1.1 Schematic representation of histone modifications. 2
1.2 Core players involved in cleavage and polyadenylation. 4

2.1 Transcription HMM model 11
2.2 Annotation of 4sU-DRBseq transcription 14
2.3 Histogram of distance Pol II front end. 15
2.4 Histogram of rough estimation of elongation rates 16
2.5 Elongation rates (control) 18
2.6 Elongation rates and time to start elongation 19
2.7 Segmented model compared to linear model 20
2.8 Comparing siRNF20 to siLacZ in segmented model 21

3.1 Illustration to the model 24
3.2 Transcript segmentation for MAVS by 2 state continuous HMM 26
3.3 28
3.4 Differences between body and tail 29
3.5 Density plots of relative read coverage 30
3.6 Significant genes in each experiment 31
3.7 Paired-end read coverage around cleavage points 31
3.8 Comparing APA database to predicted cleavage point 32

4.1 Experiments of Roadmap Epigenomics project 37
4.2 Emission parameters 38
4.3 Transitions in the inferred model 41
4.4 Spatial enrichments of states 44
4.5 Spatial enrichments of states near enhancers 45
4.6 States Comparison 45
4.7 Known annotation overlap 46
1 Scientific Introduction

1.1 Biological Background

1.1.1 Genome Organization, Chromatin and Histones

A gene is a DNA region that encodes protein or functional RNA. Genes as other genetic material are contained in genome, that is packed and organized in a complex of proteins, DNA and RNA known as chromatin, and this organization is tightly linked to specific cell type function. The chromatin is made up of nucleosomes - octamers of four different histone proteins (H3, H4, H2A, H2B) and 146 base pairs of DNA. The octamer is composed of two H3/H4 heterodimers and two H2A/H2B heterodimers. The histones (as other proteins) are subject to wide range of post-translational modifications: acetylation, methylation, phosphorylation, ubiquitylation, sumoylation, etc (see Figure 1.1 on page 2). These modifications can take place in different sites - mass spectrometry and ChIP-seq studies have identified more than 60 residues to be subject of such modifications.[19]

The most common technique to study histone modifications is ChIP-seq[3]. Different histone modifications are known to be indicative for expressed genes and for enhancers. For example [31, 26, 19]:

- H2AK9ac, H2BK5ac, H3K9ac, H3K18ac, H3K27ac, H3K36ac and H4K91ac - Usually enriched around transcription start site (TSS)
- H2BK12ac, H2BK20ac, H2BK120ac, H3K4ac, H4K5ac, H4K8ac, H4K12ac - correlates with promoter and transcribed regions of active genes
- H3K27me3 - correlates with silencing
- H3K4me1, H3K27ac - correlates with enhancers
Figure 1.1: Schematic representation of histone modifications. Adapted from Mariuswalter, Wikimedia Commons.

Histones are usually subject to multiple modifications, they are dynamic, and can appear or disappear within minutes of stimulus. There are many histone-modifying enzymes responsible for histone modifications: acetyltransferases (such as CBP/P300 and HAT1), deacetylases (SirT2) and ubiquitilases (RNF20/RNF40) etc.[19]

Other characteristic property of functional sites are highly accessible sites (DNaseI hypersensitive sites; DHS).

1.1.2 Gene Expression and Regulation

For expressing protein product, gene goes through a transcription process where DNA is transcribed to RNA resulting in mRNA, followed by translation process where ribosomes translate RNA to protein. The transcription process and following process to get to a protein product can be described in more details:

1. Transcription of the DNA to RNA: RNA Polymerase II (Pol II) recruitment to the transcription start site (TSS). The Pol II gets to promoter-proximal pausing, followed by release from the paused state, productive elongation and termination.[18]

2. Splicing and RNA processing - the mRNA goes a process of 5’ capping, 3’ cleavage and polyadenylation and splicing in which splicesome removes the introns

3. RNA export from the nucleus to the cytoplasm

The gene expression and the gene product may be subject to regulation before, during, or following each of the above steps. Recent genome-wide studies have highlighted the importance of regulation at the point of promoter proximal paused
Pol II enters into productive elongation, in addition to the regulation on Pol II recruitment at the initiation step. Another example, cell-type or condition specific expression of 3'UTR isoforms of the same gene can be result of alternative cleavage and polyadenylation.

1.1.3 Transcription and mRNA Processing

Transcription takes place from 5' to 3' direction. The premature mRNA is processed into mature mRNA, and during RNA splicing introns are removed and only exons are left. Nascent RNA constitutes only a small portion of the total RNA within cells. The mature mRNA consist of three DNA-encoded regions: The 5' untranslated region (5'-UTR), coding region, the open reading frame (ORF) and 3' UTR. As their name suggest, the 5' UTR and the 3' UTR, are not translated, but rather they are important for gene regulation.

During the elongation process RNA Pol II transcribes RNA based on DNA template. Pol II elongation rate may vary between genes, and various studies have tried to estimate it, with approaches based on Pol II ChIP, RT-PCR, cell imaging, global run-on sequencing (GRO-seq), 4suDRBseq. Those studies varies in their estimation of elongation rates, with common estimations between 1 up to 6 kilobase per minute.

1.1.4 RNA Polyadenylation

RNA polyadenylation is a molecular process of addition of poly(A) tail to the mRNA, which can contribute to the stability, nuclear export and efficient translation of mRNA. Polyadenylation signal (PAS) instructs cleavage of the mRNA in the 3'-UTR prior to addition of this poly(A) tail, and the majority of genes have multiple PASs, resulting in alternative polyadenylation (APA) sites (see Figure 1.2 on page 4), which give rise to different 3'-UTR lengths. The typical distal PAS which creates the longest 3' UTR is characterized by motif of AAUAAA. The cell type and the conditions can contribute to increase use of APA. For example during T cell activation, in which there is increase in cell proliferation may cause shortening of many 3'UTRs. Other examples include preference of 3' UTR
Chapter 1: Scientific Introduction

Figure 1.2: Core players in the cleavage and the polyadenylation process: AAUAA motif (or other motifs) and GU rich sequence recognized by cleavage and polyadenylation specificity factor (CPSF) and cleavage stimulating factor (CSTF).

It was reported that shorter 3’ UTR isoforms tends to be slightly more stable in mouse.[27]

microRNAs (miRNAs) are small non coding RNAs that can bind to mRNAs and regulate its translation and stability.[15] The binding usually takes place within the 3’UTR, where there is a site of 2-8 bases (“seed”) which is complementary to the 5’ end of miRNA.[15]

One interesting affect of the different 3’UTR lengths (e.g shortening of the 3’UTR) induced by APA is ability of the mRNA to escape from microRNAs (miRNAs) regulation due to lose of binding sites.[15] On the other hand some binding sites may become more susceptible for miRNA repression.[15].

There are various high-throughput methods to annotate and quantify cleavage and polyadenation events. Most methods are based on oligo(dT) primming off the poly(A) tail, which works well for identifying distal poly(A) sites.[27] Another method is poly(A)-position profiling by sequencing (3P-seq), based on 3P-seq tags, which can be used for identifying proximal sites and avoids false positives of A-rich regions.[27]
1.2 Computational Background

Our approach in this work is mainly **unsupervised**, e.g. the task is to infer patterns or hidden structure from unlabeled data. Unsupervised techniques include for example **clustering algorithms** such as Gaussian Mixture Models and K-means. These methods are usually used for learning individual samples, rather than complex structure such as sequences. Hence we need a more sophisticated method, which is powerful enough for learning sequences - Hidden Markov Models (HMMs). We will provide here a brief introduction to HMMs and the motivation for using them in our context.

1.2.1 Hidden Markov Model

A **Markov process** is a process in which it is enough to know the present state of the system to predict its future states. More formally let $O = (O_t, ..., O_1)$ be a sequence of observations, we can write a probabilistic model to describe it as: $P(O) = P(O_t, O_{t-1}, ..., O_1) = P(O_t | O_{t-1}, ..., O_1) \cdot \cdots \cdot P(O_1)$. By assuming **Markov property**, e.g each observation $P(O_t | O_{t-1}, ..., O_1)$ is conditionally independent of past observations $(O_1, ..., O_{t-2})$ given the previous observation $(O_{t-1})$, we can write the above as $P(O) = P(O_1) \prod_{i=2}^{t} P(O_i | O_{i-1})$. This property can be assumed for time series data or for spatial data. In the models that will further describe below, we will use it for spatial data, e.g $O_t$ and $O_{t-1}$, are two neighbor sites in the genome or a transcript.

A **Hidden Markov Model** (or simply HMM) can be described as $(S, T, E)$ where:

- **States**: $S = \{s_1, ... s_n\}$
- **Transition probabilities**: $T$ is a matrix $(n \times n)$, such that $T_{i,j} = P(x_t = s_j | x_{t-1} = s_i)$, where $x_t$ is the state at time $t$, and $x_{t-1}$ is the previous state
- **Emission probabilities**: $E = \{e_1, ... e_n\}$ is a set of $n$ probability functions assigned to each state for emitting observation, e.g $e_i(O_t) = P(O_t | x_t = s_i)$.
  - The most simple emission probabilities are for **discrete** alphabet/observations $\Sigma$: $E$ can be described as a matrix $n \times |\Sigma|$
– Emission probabilities can also describe \textbf{continuous} observations: $E$ can be described by $n \times p$ where $p$ are parameters of a probabilistic distribution. For example in case of normal distribution, $\mathcal{N}(\mu,\sigma^2)$, $p$ are the mean and the variance.

– Emission probabilities can describe \textbf{multivariate} observations either with independence or dependency between variables. For example independent Bernoulli random variables or multivariate Gaussian, respectively.

For a more detailed background on Hidden Markov Models, the readers are referred to [22, 9]. HMMs are very popular in computational biology field, as they allow to model the data with simple and understandable model. This allow us to learn the basic principles which govern the data. The parameters of HMMs (emission and transition probabilities) can be trained efficiently based on observations, and the likelihood of the data to come from a HMM can be evaluated efficiently. A trained HMM can be useful for decoding tasks (including segmentations and classifications), by inferring the most probable state path that could have generated the experimental data.
1.3 Motivation and Objectives

As there are so many types of regulations, and there are varieties of experimental procedures for studying them - either different experiments (ChIP-seq and RNA-seq) or different experiment design (measuring different timestamps), it is tempting to design and implement a generic and efficient computational framework for analyzing them. In these work we demonstrate how such framework based on Hidden Markov Models can enable simple, clear and efficient way for approaching to wide variates of biological data and experiments by studying 3 different problems:

1. Analyzing 4sU-DRB-seq data, to learn the role of monoubiquitylation of histone H2B of transcription elongation rates

2. Analyzing RNA-seq combined with various treatments (TEX, 3’ pull-down, 5’ pulldown) to provide evidence for stable, autonomous uncapped RNA fragment downstream to the coding region

3. Analyzing large dataset of ChIP-seq data from the Roadmap Epigenomics for unsupervised annotation of the genome in various cell types

The models, combined with subsequent analysis, enable us to derive interesting insights on the regulation. Here we will explain the study aims in each problem, and the experiments.
2 Histone H2B Monoubiquitylation And Elongation Rates

2.1 Motivation and Objectives

One of the histone modification which occur on the H2B histone is monoubiquitylation on lysine 123 in yeast or lysine 120 in mammals (H2BK120ub1 and hereafter H2Bub1). This modification is carried out by the E3 ubiquitin ligase Bre1 in yeast and by the heteromeric hBre1 (RNF20)/RNF40 complex in mammalian cells.[16, 34] H2Bub1 was found to correlate with various processes, such as recruitment of TFIIS transcription elongation factor.[24]

H2Bub1 is associated with promoters and coding regions of active genes,[32] and its local levels correlate with extent of gene expression and transcription elongation rate.[11] Evidences for the link between H2Bub1 and RNA Pol II elongation rates include low levels in exonic region[11] where Pol II frequently pauses.[20]

Transient RNF20 knockdown in HeLa cells leads to a reduction in overall H2Bub1 levels, but impacts the steady state level of only minority of transcripts, while the majority of transcribed genes are not significantly affected.[25]

Although H2Bub1 correlates with elongation rates, it is not clear whether the correlation is due to causative role of H2Bub1 in regulating elongation rate or vice versa, and there is no direct in vivo evidence that H2Bub1 is functionally needed for transcription elongation in mammalian cells.

In this study our aim is to directly test the role of H2Bub1 in transcription elongation, by measuring the effect of RNF20 knockdown and H2Bub1 down-regulation on the elongation rates of thousands of genes.
2.1.1 Experimental Schema

The lab experiments were done by Gilad Fuchs from Moshe Oren’s lab in Weizmann Institute of Science.

For measuring elongation rates one can use global run-on sequencing (GRO-seq) [8] or 4suDRBseq [11], both methods enable genome-wide measurement of transcription elongation rates. In 4suDRBseq:

1. Cells are treated with 5,6-dichlorobenzimidazole 1-β-d-ribofuranoside (DRB), which reversibly blocks transcription and release of Pol II from pausing, in vivo for 3 hours.
2. DRB is removed to re-initiate transcription and nascent RNA is labeled by 4-thiouridine (4sU) for few minutes.
3. RNA is isolated, 4sU-tagged RNA is biotinylated, purified on streptavidin beads, and sequenced using RNA-seq.

Repeating the above procedure for multiple time points after removing of DRB (0, 8 and 12 minutes), can be used for genome-wide estimation of the elongation rates of genes using computational model. (see Figure 2.2 on page 14)

Here we are evaluating the elongation rate for genes in cells transfected with different siRNA oligonucleotides:

- **siRNF20** - cells with siRNA directed against RNF20 (RNF20-knockdown; H2Bub1 depletion)
- **siLacZ** - cells with siRNA directed against lacZ (control)

By comparing the elongation rates in the different conditions, we will try to learn the role of H2B-ub1 in transcription elongation. We use 2 replications (384, 385). Additional data in this study include Western blot and mRNA-seq.

The data is available in GEO as GSE69738.

2.1.2 Computational Model

Based on our understanding of the transcription process, we can model the data as follows: After removing DRB, for each gene it takes $t_0$ for the first Pol II to...
2.1 Motivation and Objectives

Figure 2.1: Simple transcription HMM model. Each window of size 50bp is discretized based on percentile (5, 50, 75, 90). A simple two states Hidden Markov Model, transcribed state and non-transcribed state model the data. The transcribed state have higher probabilities for emitting high read count, and the non transcribed have high probablitiy to emit low read counts. There is 0 probability for transition from non transcribed state to transcribed state.

start elongation. As the time goes by, other molecules of Pol II start to elongate the transcript. This result in higher read counts near the TSS, and the read counts get lower as we advance along the transcript (in the direction of 5’ to the 3’). The read counts are getting more or less to plateau of zero, in the positions along the transcript which follows the last position where the first Pol II has reached to when the RNA was harvested.

Based on this process we will analyze the data in two steps:

1. **Elongation region annotation and determining Pol II front end** - discriminate between transcribed and non-transcribed regions, and infer the position where the Pol II stopped elongation. This can be achieved using simple HMM (see figure 2.1, source code: [23]). In this model we have 2 states:
   - **Transcribed state** - Usually it emits many reads and sometimes it may emit few reads or nothing.
   - **Non-transcribed state** - as opposite to the transcribed state, it emits reads with small probability (“noise”).
The automata works as follows: It starts in the transcription start site (TSS) with transcribed state, it emits a read count, and then it go to the next position in the transcript and with some probability moves to non transcribed state.

Using annotated regions of elongation after time $t$ we can derive rough estimation for the elongation rate. However, such model is insensitive to the read counts and does not take into account the general pattern of decrease in read count.

2. **Elongation rate estimation model (“Triangle”)** - Let $t$ be the time of the experiment, and let $t_0$ (release time) be the time for starting productive elongation and let $e$ be the elongation rate. We will define $t_{\text{eff}} = t - t_0$ as the effective time of elongation. We would expect the read counts along the transcript to align to a linear line which will reach $e \cdot t_{\text{eff}}$, and plateau of zero reads afterwards. To train the model, we will estimate the time to reach to distance $x$ from the TSS by $t(x) = \frac{x}{e}$. The expected time for which the position was reachable to Pol II is then $t_{\text{eff}} - t(x)$, and we can use it to fit the read counts. The model will be evaluated based on the average of two objectives:

   a) The RMSE of the predicted read count $\hat{Y}$ and the actual read count $Y$
   
   b) The difference between the HMM elongated region and $e \cdot t_{\text{eff}}$.

Such model takes advantage of the triangle like shape of the read counts on the one hand, and on the other hand is not too sensitive to noise in read count by taking the elongation estimation of the HMM.

3. **2-steps elongation model** - As it may be oversimplification to assume the Pol II elongation rate is constant and uniform, we would like to evaluate a more realistic model, and suggest a model with 2 phases of elongation:

   a) Pol II starts the first phase of elongation with elongation rate $e_1$

   b) In a transition point, $x_0$, it changes its elongation rate to $e_2$

An evidence for such model can be seen from the data - as some transcripts show a pattern of different read coverage along the transcript. Similar to the previous model, $t_0$ will denote the time for starting productive elongation. Now instead of $e$, elongation rate, we will assume $e_1$ elongation rate from TSS to
position $x_0$, and from there on elongation rate is $e_2$. With this model, distance $x$ from TSS is reachable for Pol II as: $t(x) = \begin{cases} \frac{x}{e_1} & x \leq x_0 \\ \frac{x}{e_1} + \frac{x-x_0}{e_2} & x > x_0 \end{cases}$. Similar to above, we will fit $t_{eff} - t(x)$ to the read counts. As this model is an extension of the Triangle model, we should make sure it does not over-fit. To justify this model we will use f-score to compare it to the reduced model (2).

The HMM was trained using Baum-Welsch, and the subsequent models were trained using energy based optimization algorithm basin-hooping as implemented in scipy library.[30] The analysis was done on windows of size of 50bp. Only genes of length greater than 35,000 were considered to get reliable estimation of elongation.

Analysis of RNA-seq data can be complicated as a single gene can have multiple transcripts, with different exon usage. Hence reads originated from the same position in the genome can be attributed to different transcripts. In the context of 4sU-seq, reads originated from exons may due to contamination from mature RNAs. To avoid those obstacles, we excluded exonic 4sU-seq reads, and transcripts originated from same transcript cluster (based on known isoforms) were analyzed together to take into account various possibilities of annotated exons.

## 2.2 Results

### 2.2.1 Genome-Wide Annotation of Elongated Regions in 4sU-DRBseq

As the first step to analyze the 4sU-seq data we would like to annotate the Pol II elongated region at time $t$ for each transcript using HMM.

Figure 2.2 shows an example of the 4sU-DRBseq data and the derived annotation. As we can see 4sU-DRBseq is noisy experiment, and there could be read counts higher than zero even in position which are clearly not part of the transcript. The HMM approach annotated transcribed regions, and gives a good annotation to the transcription.

As expected, transcripts get longer in the experiments of 12 minutes Pol II elongation compared 8 minutes - this is clear both in the annotation and the raw data, as
Figure 2.2: A screen shot from UCSC Genome Browser showing the 4sU-DRBseq for UBAP2 gene, chr9:33,921,691-34,048,947. The experiments include 2 set of replicates (384, 385), with siLacZ as control (“L” tracks) and siRNF20 (“R” tracks). Arrowed tracks below the 4su-DRBseq shows the inferred elongation regions based on the HMM.

well as in histogram (see Figure 2.3) in which we compare the difference between elongated region at 12 minutes and 8 minutes based on the HMM annotation. From the histogram we can note that for some transcripts (less than 15%), the elongated region is no longer, and sometimes even shorter in 12 minutes compared to 8 minutes. Such situations can appear due to wrong annotation - in low expressed genes where the signal to noise ratio is small, it is hard to discriminate between transcription and no transcription. In addition, different transcripts may overlap, may affect the annotation of the other. The correct annotation of transcription is not always clear. We ignored such transcripts in the further steps of the analysis.

From the HMM annotation we can derive rough estimation of the elongation rates: \( \frac{D}{t} \), where \( D \) is the distance of elongation until time \( t \). This can be done either
2.2 Results

Figure 2.3: Histogram of distance between elongated region in transcript after 12 minutes and the same transcripts after 8 minutes in the control (siLacZ). Only transcripts with elongation length longer than 1.5kb in both 12 minutes and 8 minutes are included, to account only elongating transcripts. The first cluster (around 0) indicate transcripts for which the elongation annotation is invalid resulting in negative, zero or very slow transcription rate, which may explained due to high noise. For some transcripts (<15%), the elongated region is no longer, and sometimes even shorter in 12 minutes compared to 8 minute - where the annotation is unreliable and is ignored in further analysis.

Based on the elongated regions of each of the experiments, or based on the difference between them. As we can see in figure 2.4, most transcripts have elongation rates of between 1.5kb/min to 4kb/min, which is in the range similar to previous studies. However, such estimation does not take into account the read counts along the transcript, and wrong annotation can easily lead to wrong estimation of elongation rate. In addition it assume constant elongation along the transcript.

Based on this rough estimation we can derive some interesting insights. First we note that histogram of elongation rates estimated based on 12 minutes time point (red) is right shifted compared to 8 minutes (green), with slightly faster elongation rates. Furthermore, the elongation rates are even faster based on the difference between the experiments $\Delta(12 \text{min}, 8 \text{min})$ (blue). This may indicate there is a pausing in
the Pol II or small delay in the Pol II release time, which is canceled in this estimation. Estimation of elongation rates based on longer experiment (12 minutes) have slightly lower coefficient of variation ($\frac{\sigma}{\mu}_{12} = 0.447$, $\frac{\sigma}{\mu}_{8} = 0.466$).

Based on those insights, we will try to derive a more reliable elongation rates estimations - with a model that takes into account also the read counts along the transcript, and models the release time.

### 2.2.2 Estimation of Elongation Rates and Release Time

We fitted the read counts of the elongated region to a linear model (see Computational Model), to derive elongation rate ($e$) and release time ($t_0$). The same procedure was used for both control samples (siLacZ) and H2B-ub1 depleted samples (siRNF20).
For most of the transcripts, elongation rates were found to be between 1kb/m to 4kb/m (see Figure 2.5). As for elongation start time of the first Pol II, we got bi-modal distribution - where one group of transcripts have small or zero $t_0$, and the second group have $t_0$ of around 2 minutes (which was the higher limit by constraint of $t_0 \leq 2$). The $t_0$ distribution can be explained by two groups of transcripts: Small group for which the elongation start following the removal of DRB (small $t_0$), and a second group for which it took around 2 minutes for starting elongation. A different explanation is that for shorter transcripts the model tended to prefer $t_0$ of around 0, while for longer elongating transcripts we have more reliable estimation (~2min).
Figure 2.5: Elongation rates (control). Histogram of elongation rates of transcripts in control (siLacZ) based on the linear model. The histogram shows the elongation rates for control (siLacZ) for 8937 transcripts. Sample transcripts: BRCA1 (chr17:41,232,955-41,277,500) slow elongated gene (1.04 kb/m), LGMN (chr14:93,170,152-93,215,047) medium-rate elongated gene (E=2.2 kb/m) and FBXW11 (chr5:171,389,396-171,433,941) as fast elongated gene (3.5 kb/m).

As for the difference between siRNF20 to control, there is a slight reduction in the elongation rates of transcripts in siRNF20 compared to siLacZ (see Figure 2.6). Interestingly, the main difference between H2Bub1 depleted cells and the control was in the release time, e.g the time to start productive elongation.
2.2 Results

Here, the time to start elongation for siRNF20 was longer. This finding indicate H2Bub1 is important mainly for the start of elongation, and it is not rate limiting factor for elongation rates.

![Figure 2.6: Elongation rates and time to start elongation in linear model.](image)

On the left: cumulative distribution of elongation rates ($e$) of same set of transcripts for siLacZ and siRNF20 based on the linear model. On the right: cumulative distribution of release time ($t_0$).

2.2.3 Segmented Model Estimation of Elongation Rates

While the linear model provides a reasonable estimation of elongation rates and release time, for many transcripts the read counts is not distributed on a linear line, which suggest there is more than one phase of elongation and the model can not explain it well (see for example figure 2.7). Hence we will evaluate 2-steps elongation model (see Computational Model).

As this model extends of the previous model (the linear model have $e, t_0$ parameters, the segmented model has $e_1, e_2, t_0, x_0$), we would like evaluate the improvement in its predictive power (in terms of RMSE) to the linear mode described above. For 40% of the transcripts in the siLacZ, and 49% of the transcripts in siRNF20 the segmented model shows significant improvement compared to the linear model ($p < 0.01$ based on t-test). This suggest that modeling only a single constant elongation rate for each transcript is too simplistic.

Based on this segmented model, $t_0, x_0, e_2$ have different distribution in H2Bub1
Figure 2.7: Segmented model compared to linear model. PPHLN1 gene (transcript name uc001rnd.3; chr12:42,719,946-42,835,765). Top: Control (siLacZ), Bottom: H2Bub1 depleted (siRNF20). Left: 8 minutes, Right: 12 minutes. Red line - shows the read count along the transcript from TSS. Horizontal gray line show the Pol II front end (e.g elongated region) based on HMM. The two phases elongation is most clear in the 8 minutes samples (both) based on the observed data and the difference between the segmented model (green) and linear model (dashed blue).

For siRNF20 58% of the transcripts have equal or faster elongation rates than siLacZ (see Figure 2.8). For the release time, 38% of the transcripts have longer or equal release time in siRNF20 compared to siLacZ.

With just two replications of the experiments (384, 385), it is hard to give confidence for the model predictions. However, with consistent computational model for both siLacZ and siRNF20, we cannot observe significant clear differences in the elongation rates. The differences in \( t_0 \) and \( x_0 \) may suggest H2b-ub1 play a role in Pol II recruitment or regulation of proximal Pol II pausing in gene specific manner.
2.2 Results

Figure 2.8: Comparing siRNF20 to siLacZ in segmented model. The histograms show that difference in the segmented model parameters for transcripts in H2Bub1 depleted cells (siRNF20) compared to control (siLacZ). Top left: shows the difference in the **first elongation rate** ($e_1$). Top right: difference in the **second phase elongation rate** ($e_2$). Note that histogram is right shifted (slightly faster elongation rate in RNF20). Bottom left: shows difference in **release time** ($t_0$). Bottom right: shows the difference in **position of change in elongation rate**.
3 Evidences for Stable, Autonomous Uncapped RNA Fragments

3.1 Motivation and Objectives

mRNA molecules undergoes various modifications which play important roles in regulation. One of the common modifications is alternative polyadenylation - cleavage and addition of poly(A) tail to the 3’ end of the shortened RNA molecule. The process can give rise to different transcripts originated from same gene.

In this project we set out to explore the conjecture that the cleaved part during the shortening result in a stable 3’UTR RNA tail which lives alongside the transcript of origin. This conjecture may have interesting aspects in terms of miRNAs regulation. Such mechanism suggest that in addition of lose of miRNAs binding sites downstream to the APA site, the cleaved part may still have a functional importance as competitive inhibitor to the miRNA - as it may be a potent target for miRNAs.
Figure 3.1: Illustration of the suggested model. The first 3 steps are based on the known model of transcription and related RNA processing (5’ capping and polyadenylation). The model suggest that the cleaved part (green) is stable and could potentially play a functional role as inhibitor of miRNAs. Adaptation from Kelvinsong, Wikimedia Commons (CC-By-SA 3.0) https://commons.wikimedia.org/wiki/File:MRNA.svg

3.1.1 Experimental Schema

The lab experiments were designed and done by Yuval Malka in Michal Berger’s lab. Cells were subject to the following treatments followed by RNA-seq:

- **Untreated** (Control) - regular RNA-seq
- **TEX** (Terminator 5’ phosphate-dependent exonuclease) - Degrades mRNAs
3.1 Motivation and Objectives

that are not 5’ capped

- **5’-pulldown** or **CAP IP** - immunoprecipitated (IP) using anti-CAP antibody
- **3’-pulldown** - Biotinylated guanosine triphosphate (GTP) followed by pull-down of the newly capped fragments using streptavidin beads, which enrich the tails fragments

The experiments were done in U2OS cells and HEK293 cells. In HEK293 cells, additional similar experiments were done, but with addition of miRNAs. Most of the analysis to this part of the experiment, and in particular the miRNAs affects were done by Avital Steiman Shimony from Hana Margalit lab.

Another set of experiments is incubation in α-amanitin and TEX, followed by 3’-sequencing, to provide evidence that the process takes place as a post transcriptional.

The data is available in GEO as GSE84068\(^1\).

### 3.1.2 Computational Model And Analysis

Based on the above lab experiments, and the resulted RNA-seq we would like to evaluate the conjecture.

Analysis was performed separately for each transcript and for each treatment (TEX, 3’-pulldown, CAP IP).

First, the average read coverage for both treatment and control was calculated for overlapping 50bp windows (in 20bp offsets) along the coding regions and 3’UTR (excluding introns). Coverage values were then transformed to log scale. (see figure 3.2)

We considered all windows that overlap coding regions (from all transcripts), and fitted a linear regression model \( Y = a \cdot X + b \) for each of the treatments, where \( Y \) is the log-transformed window coverage for every treatment, and \( X \) is the log-transformed window coverage for the control. This allowed us to overcome experiment-specific bias and infer the offset and ratio between mean read coverage (per window) in every treated condition and control. We then applied the linear model to yield the

\(^1\)http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE84068
Figure 3.2: Transcript segmentation for MAVS by 2 state continuous HMM. To segment transcripts body and tail fragment we use 2-state HMM with emission based on normal distribution. The state emission explains the read counts in the exonic windows along the transcript. The figure shows a schematic view of MAVS gene (introns excluded), the read counts for TEX and control (untreated) cells, and the maximum likelihood segmentation of the model.

expected read coverage for each transcript along both coding and 3' UTR windows, and deviations from the regression model were calculated.

Next, we trained a Hidden Markov Model (HMM), with two internal states corresponding to “Body” and “Tail” regions. Each state was characterized with a Gaussian emission of continuous variables. The parameters of the model were optimized using the Baum-Welch algorithm. Transition from the “Body” to the “Tail” state were not allowed. Finally, the most probable Body / Tail cleavage point was identified for every transcript / treatment as the maximum likelihood transition point between the Body and Tail states, using the Viterbi algorithm.
3.2 Results

3.2.1 Comparing Coding Region to 3’UTR

By inspecting TEX treated and regular RNA-seq in U2OS cells, we note that for many genes there is higher TEX/Control read count ratio in the coding region compared to the 3’UTR (for example GPD2; see Figure 3.3). To determine whether it is a wide phenomena and whether it is significant, our first validation was genome wide search for transcripts for which the average read count in the 3’UTR region is significantly lower compared to coding region.

Our null hypothesis is that a transcript is composed of one fragment. Under the assumption that non 5’-capped RNA are negligible, this would imply that the distribution of read count along a transcript would be more or less similar in untreated and the TEX treated cells. Hence the read counts ratio (TEX/Control) in exons should follow normal distribution (see Figure 3.3).

A similar analysis was done also for the 5’-pulldown and for the 3’-pulldown (for which we look for significant higher ratio). Out of 54,469 annotated transcripts: 23,379 transcripts are enriched in coding region compared to 3’UTR in TEX and 18,157 are enriched in 5’ pulldown, and 33,041 are enriched based on the 3’ pulldown. This analysis indicate that for many transcripts the treatments (TEX, 5’ pulldown and 3’ pulldown) effectively allows to distinguish between different segments along the transcripts which undergoes processing.

In the this analysis we compared the coding region to 3’UTR. However the actual cleavage point in which the RNA-seq/TEX ratio along the transcript starts to diverge to two distributions may be within the 3’UTR, so we would need a more fine grained approach.

3.2.2 Genome Wide Annotation of Cleavage Points

To determine cleavage point from which there is difference in the RNA-seq read coverage for TEX treated compared to untreated cells (see Figure 3.2 on page 26), we used continuous (Gaussian) HMM, based on observations of read coverage of
RNA-seq along the transcript. We then focused on transcripts in which the cleavage points fall downstream to the coding end site.

By breaking the transcript to A and B segments (body and tail), upstream and downstream to the cleavage point respectively, we note the difference in the read coverage in TEX treated compared to control (see Figure 3.4). This shows clearly that for many transcripts, the distribution of read counts in body segment is different between the TEX and the untreated cells. The histogram of the ratio of the TEX/Untreated read counts in tail is left shifted, indicating less reads in the tail compared to body, providing again evidences for the effectiveness of TEX method to enrich body segments.
3.2 Results

Figure 3.4: Differences between body and tail. **Upper left** - TEX compared to control in the A (body) segment. A black line shows a local regression model fitting the read coverage of the TEX treated based on the control. **Upper right** - TEX compared to control in the B (tail) segment. A black line indicate the same local regression model. **Bottom** - Comparing the expected TEX according to the control to the actual TEX. We note that the TEX treated is lower than what we would expect in the B segment.

Out of 54,383 annotated transcripts, the HMM found 12,578 transcripts with clear difference in body/tail in TEX treated, 11,108 transcripts in 5’ pulldown (CAP) treatment, and 14,589 transcripts in 3’PD treatment (see Figure 3.5). As a more careful estimation to the scope of this process, we looked for the overlap of genes reported in the 3 experiments (see left figure of 3.6). Overall, in all three treatments we report 2123 common genes showing a significant coordinated change in read coverage, with an average 32.7% decrease in tail vs body coverage for TEX, 35.3% increase in tail vs body coverage for CAP IP, and 40.8% decrease in tails for the 3’-pulldown experiment. These results suggest that for thousands of genes, stable cleaved transcript tails are present in the cell in substantial amounts.

To validate whether the results are reproducible and whether their scope is not cell type specific, we compared similar analysis for U2OS cells and 293HEK...
Figure 3.5: Density plots of relative read coverage. Shown are density plots comparing the relative read coverage before and after treatment, for bodies (x-axis) and tails (y-axis) for TEX and 3’PD. Red: density of transcripts predicted to be cleaved be the HMM, Blue: transcripts for which there is no prediction for cleavage.

**cells** (see right figure 3.6). There is an overlap between the groups, with **3,432 significant genes found in both cells**.

As a further validation for the computational model and the suggested segmentation, the RNA-seq read coverage for paired-end untreated cells were evaluated in the region surrounding the cleavage point (see Figure 3.7). The breakpoint is characterized by low read coverage compared to the surrounding area even only based on the untreated cells RNA-seq, giving a further support for the segmentation.
3.2 Results

**Figure 3.6:** Significant genes in each experiment. **Left:** Venn diagram of genes with predicated cleavage point in the 3’UTR, with significant (p<0.01 after FDR correction) difference in read coverage on the body compared to tail, in each of the 3 treatments of TEX, 5’ prime, 3’ prime. Of around 17400 genes, 6068 (34.9%) genes were found significant in TEX, 5222 (29.95%) were found significant in 5’ selection, 6501 (54.41%) were found significant in 3’ selection. **Right:** 6,068 genes (34.89% of the genes) in U2OS and 6578 genes (37.55%) in HEK293 had a predicated cleavage point with significant difference in the read coverage between body and tail, with overlap of 3432 genes.

**Figure 3.7:** Paired-end read coverage around cleavage points. Average read count around break point as predicated by the HMM based on the TEX vs control (untreated) data. The red line shows the paired-end read coverage, which decrease around the cleavage point, and the blue line shows the exon coverage as a baseline.
3.2.3 Characterization of Cleavage Point

With those promising results, we set out to explore whether the cleavage process is associated with the alternative polyadenylation process.

First, we compared our results to a database of APA sites that were previously described[12]. If the cleavage suggested by our model is related to alternative polyadenylation, we would expect to see spatial correlation between the predicted cleavage point to known APA sites. Though the previously described APA sites data derived from various cell types, we did see good spatial correlation with our predicted model (see 3.8). However, we did not find known motifs related to alternative polyadenylation (such as AAUAA).

**Figure 3.8:** Comparing APA database to predicted cleavage point. Location of cleavage point as inferred by the HMM using data TEX and control of U2OS, compared to APA database. Excluding transcripts where the predicted cleavage is not in 3’UTR.

To further investigate our hypothesis with more conservative method for characterization of polyA sites from same cell type, we analyzed data derived from 3’-end
3.2 Results

**RNA-seq.** The cleaved model suggest there should be two peaks: proximal peak for the shortened transcript, and distal peak - which could originate either from the canonical (full length) transcript, or from the tail segment.

In TEX treated cells we observed decrease in the distal peak, as expected from the model as TEX degrades non capped (tail) fragments. The average decrease was **22% in the distal compared to proximal peak heights.**

To characterize whether the process occurs post-transcriptionally, the above experiment was repeated, but now with α-amanitin, an RNA polymerase II and III specific inhibitor. Transcriptome-wide analysis of the 3’-end RNA-seq data showed a significant decrease (p<1.8e-17) in the distal peak height compared to the proximal peak following TEX treatment in α-amanitin treated cells, supporting the hypothesis of post-transcriptional cleavage.

We compared the location of the putative cleavage site (as predicted by the HMM model following TEX) with the nearest 3’-end RNA-seq peak (APA site) in the same cell type (U2OS). For 80% of the cleaved transcripts, the predicted cleavage site appeared within 100nt from the most proximal 3’-end RNA-seq peak. These results may suggest that the cleavage point occurs at APA sites.
4 Epigenomic Annotator

We will now turn to visit an interesting challenge of unsupervised annotation of the genome based on ChIP-seq dataset in many cell-types. As we would like to study the chromatin modification orchestra, we would build a multi-dimensional model, taking into account multiple experiments of different chromatin modifications.

4.1 Motivation and Objectives

Thanks to recent years technological advances in the sequencing technology, massive amounts of data is generated everyday in labs. The new epigenomic data is becoming publicly available as part of large scale epigenomic initiatives e.g. Roadmap Epigenomics[4], ENCODE[7] and Mouse Encode[28]. These studies provide thousands of publicly available dataset of ChIP-seq data (and other types of data), and invaluable opportunity for studying the regulation in the chromatin level and to discover functional sites. With the ever increasing big-data and large scale epigenomics projects, there is also increasing need for new methods for integrative analysis of data.

As ChIP-seq is quite noisy experiment, one of the common first steps in analyzing ChIP-seq data is peak-calling. Peak-calling aims to discriminate between real peaks and noise, and there are various peakcallers, such as the popular MACS[33] and HOMER[13]. Methods for integrative analysis of ChIP-seq data usually combine the data derived from various experiments after peak calling. ChromHMM[14], one of the most popular of such tools, gets as input peaks, and it uses multivariate Hidden Markov Model, where the emission distribution is a product of independent Bernoulli distributions. Other methods such as treeHMM[5] takes advantage of cell type hierarchy but also requires binarized peaks.
In this project we develop **Epigenomic annotator**. Epigenomic annotator is a probabilistic framework based on Hidden Markov Models for integrative analysis of large scale epigenomic data. Here we demonstrate its power by analysis of ChIP-seq data from Roadmap epigenomics project[4]. We challenge the integrative analysis based on peak-calling procedure by studying the data (after preprocessing of ranking and standardization, see below) directly as coming from multivariate Gaussian distributions. By avoiding peak calling and using standardized data, our method is more sensitive to the signal intensity and allow more sensitive annotation of the genome, providing new interesting overview of the epigenome. As a demonstration to its efficiency we challenge our method with a task to annotate more than 100 cell types from Roadmap Epigenomics project.

### 4.1.1 Experimental Schema and Data

Publicly available data collected in the Roadmap Epigenomics[4] project was downloaded from National Center Of Biotechnology (NCBI)[1]. As our method aim to study a general model for different cell types and different read depth, it is important to standardized the data.

4.1 shows a eye-bird overview of the dataset, with percentiles values of some high throughput ChIP-seqs experiments.

The read coverage or the level of signal for different cells and experiments may vary based on the sensitivity of experiments and due to data derived from cell population rather than single cell. Some types of data may present interesting signal below 95% percentiles (in particular H3K36me3 which correlate with transcribed regions). Such data could be lost in binarization based on threshold, and it may be interesting to take such lower signal into account when we model epigenome. This may be especially interesting for broad peaks where incorrect threshold may result in splitting a peak to multiple peaks and having a model with less assumptions on the peaks patterns. However, this could be a challenging task, as even with read depth normalization, similar experiments on the same cell type may have different sensitivity and are not always directly comparable (see 4.1). Hence it is important to scale/normalize the data and to have comparable experiments. Here we rank the data, and use simple standardization by $\mu = 0$ and $\sigma^2 = 1$ of the ranked data.
4.1 Motivation and Objectives

4.1.2 Computational Model And Analysis

We model the chromatin modifications as coming from multiple states, emitting standardized read counts in multiple dimensions, each dimension for a different chromatin modification. We use multivariate Gaussian as the emission distribution of the different states, which gives an intuitive interpretation in multivariate view. To provide unbiased model, we do not set any assumption on the state transitions, hence having a fully connected architecture and the model is fully inferred from the data.

We train a combined model based on data derived from multiple cell types to have a general and comparable annotation of genome. We then use the trained model to annotate or classify the genome as coming from multiple states. To learn functional
importance of the states we compare the enrichment of existing annotations on different states.

The model was trained based on 5 ChIP-seqs of chromatin modifications: H3K27me3, H3K36me3, H3K4me1, H3K4me3, H3K9me3. The decision to use those modifications rather than other is based on the availability of such data (107 different cells types) and the availability of similar annotations based on such dataset (ChromHMM) which would allow us to easily compare the resulted model to an existing method.

4.2 Results

4.2.1 Inferred Model: 15 States Model Of The Chromatin

We will examine the results of training the 15-states fully connected, where the emission of the states is derived from multivariate Gaussian distribution, and inspect the learned parameters (emission and transitions).

Figure 4.2 presents the mean emission of the states.

Some of the interesting states include:

- **TSS (N)** - this state is characterized by high read coverage for H3K4me3 known for TSS, and H3K4me1 which is expected in TSS and enhancers

- **Main transcription states**: Tx High (K), Tx Med (G) and Tx Low (C) are characterized by H3K36me3, common in actively transcribed regions, where K

![Figure 4.2: Emission parameters. A heat-map showing the mean emission (in ranked normalized values) of histone modification of the histone states.](image)
have very high enrichment ($\mu = 2.53$) to this modification (with mean emission of 1.69 and -0.16 for G and C respectively).

- Of a note, Bivalent (O) and Flank TSS (M) are also characterized with high H3K36me3 (0.89 and 0.17, respectively) but are also enriched with other modifications (H3K27me3 for O and H3K4me1 for M).

- Flank histone states: Pre Flank TSS (I) and Flank TSS/Enh (M) are characterized by H3K4me1 associated with flank histones.
  - Flank TSS/Enh (M) have higher emission for H3K4me1 compared to Pre Flank TSS (I), and is enriched near TSS. It is also associated with H3K36me3 indicating active transcription.
  - Enhancer (L) is also associated with high H3K4me1, but also have higher emission for H3K4me3 (associated with active TSS).

- Wide repression (H) state is common in repressed chromatin, where wide repression is associated with H3K27me3.

- Anti-enhancer (J) and Narrow repression (E) are both associated with high H3K9me3 known for heterochromatin and repressed genes. Those states marks marks shorter regions of repressed chromatin compared to H (evident also based on the self transition probabilities). In some cell types, such as pooled leukopaks T cells, J plays role as “background state” (with around 30% of the chromatin marked with this state)

- Bivalent (O) is characterized by modifications associated with both repressed chromatin such as high H3K27me3 and H3K9me3 and active chromatin H3K36me3 and H3K4me1.

- N/A (A) state or no (unique) alignable reads state is a state with zero or very low read coverage of all type of modifications. It is technical state, and regions associated with this state are characterized by low mappability (average unique mappable DNA there is less than 10%).

- Background (D) state is the common state in most cell types (hence is name) is associated with facilitative (non constitutive) hetrochromatin (H3K27me3).
As for the transition parameters while the model is fully connected, allowing transition from any state to any state, we see that some transitions are less likely than others (see Figure 4.3). In general we can note few different groups of states:

- **Transcription states** - Tx low (C), Tx Med (G), Tx High (K) which all are associated with H3K36me3. The states are highly likely to have transitions between each other.

- **Regulation states** - Pre Flank TSS (I), Flank TSS/Enh (M), Enhancer (L), TSS (N) states with high emission probability for methylation of H3K4.

- **Repressed/inactive chromatin** - states associated with H3K9me3, where Anti-enhancer (J), Activation Repression (F), Bivalent (O) are also associated with H3K27me3

- **“Background” states** - Background (D) and N/A (A)
Figure 4.3: Transitions in the inferred model. The trained transition parameters as a graph. Transitions with weight of less than 5% shown as dotted lines. There are a clear groups of highly related transcription states (Tx states), active regulation states (TSS, Flank Tss/Enh, Enhancer, Pre Flank TSS), and repression states.
4.2.2 Functional Meaningful Annotation

To verify our annotation gives functional meaningful results we would like to compare the derived annotation to existing known annotations. We will examine the spatial enrichment of each of the states around UCSC genes (see heat-maps in figure 4.4).

The enrichment heat-map shows the inferred model catch functional meaningful results. We can see TSS is strongly enriched for TSS state as we could expect based on previous studies. Over 43% of the transcription start sites of known genes in UCSC are annotated with TSS state in adipose derived mesenchymal stem cells. The surrounding of TSS is enriched by Enhancer state (H3K4me1 and H3K4me3), with more than 23% of the genes are annotated with enhancer state around 1k upstream the TSS. Downstream to the TSS, there is enrichment for Flank TSS/Enh state (H3K4me1), with maximum enrichment around 2k downstream to the TSS with more than 15% of the genes have this annotation. Additionally background is enriched for intergenic regions. Transcription states (Tx Low, Tx Med, Tx High) are enriched within the gene bodies.

Comparing the derived annotation from fetal brain cell to experimentally verified brain enhancers from VISTA enhancers[29] (see 4.5), shows there is enrichment for Pre flank TSS, Enhancer, Flank TSS/Enh, Bivalent states around enhancers, with 7%, 19%, 20% and 11% (respectively) of the verified brain enhancers annotated accordingly based on data derived from fetal brain (week 17). Of the verified enhancers, 36% are classified with the background state.

4.2.3 Comparing to ChromHMM

We now turn to examine the method compared to a similar popular tool - ChromHMM with similar 15-states model.

First we compared the annotation as defined by both programs for same input experiments (Foreskin Fibroplast primary cells). For ChromHMM, almost 54.5% of the genome is annotated as quiescence state (state 15), while in our segmentation the best match is background state which match also repressed polycomb (state 14) which covers 36% of the genome.
4.2 Results

As the methods are based on different emission functions there is no clear one to one mapping. However, for ease of comparison, we used the Hungarian method to find the best match between states, according to overlapping annotations (see Figure 4.6). There are groups of similar states: 1._TssA (H3K4me3 enriched) and 2._TssAFlnk (H3K4me3+H3K4me1) are similar to TSS and Enhancer, respectively. Another group of related states are 13_ReprPC, 14_ReprPCWk to Bivalent and Activation repression. However, some states may have multiple good matching - Flank Tss/Enh state in our method is associated mainly with 7._Enh, but is also associated with 5._TxWk state.

Next we compared the classification to UCSC annotations (see 4.7). **72% of the positions classified as TSS state are known to be TSS**, while in ChromHMM **68% and 64% of the TssA (H3K4me3) and TssBiv states (H3K4me3 and H3K27me3) are known to be TSS.** Assuming TSS are well described in reference UCSC annotation, we suggest this model is slightly more meaningful as it better agrees with reference annotation. Additionally for transcription states in our model are more enriched in introns compared to ChromHMM (82%, 85% and 84% of Tx Hi, Tx Med and Tx Low respectively, compared to 85% and 76% in 4_Tx and 3_TxFlnk).

As for transcription states, gene enriched states (Tx Low, Tx Med, Tx High) in our model were compared to Tx and TxWk in ChromHMM. For this comparison we used sample Foreskin Fibroblast Primary Cell, for which there is RNA-seq data available (GSM751277) which is assumed to provide a ground truth for transcription. In the suggested model 10.7% of genome is classified in transcription states while in ChromHMM model 16.13% of the genome is classified in transcription states. However, when evaluating the RNA-seq read counts in regions classified as transcription (see Figure 4.8), we note that **10% of the transcribed states in our model have no read count RNA-seq**, compared to **18.7%** in ChromHMM. This indicate our model may perform better in classification for transcription states, as it provide less false positives and overall the read counts are higher.

Another interesting aspect to compare the methods could be by examining eye-bird distribution of states across different cell types. As different cell types have different gene expression patterns, they could have different distributions (see cluster-map 4.9). We note that in both methods the chromatin states are not equally distributed in all cell types, as some states are more common in specific cells than others.
Figure 4.4: Enrichment heat-map of different states near known TSS and TES (based on UCSC annotations) for adipose derived mesenchymal stem cells. A very strong enrichment for TSS and medium enrichment to enhancer is clearly seen around TSS. Around the TES, we can clearly see the transition from high enrichment of transcription related states (e.g. Tx Med and Tx High) upstream to the TES compared to their enrichment downstream to TES.
Figure 4.5: Enrichment heatmap of states near verified VISTA brain enhancers. There is an enrichment for Enhancer (L), Flank TSS/Enh (M) around enhancers. Anti-enhancer is displaced from enhancers sites, but may appear around them.

Figure 4.6: States comparison: Heatmap showing the number of segments assigned to each state (rows - ChromHMM states, columns - our method).
### Chapter 4: Epigenomic Annotator

#### Figure 4.7: Known annotation overlap. Top: The distribution of each annotation over states in our annotation (left) and in ChromHMM (right). 25% of the exons are annotated as Tx High, and 22% of the known TSS are annotated as such based on our model. In ChromHMM, 21% and 19% of the exons are annotated as 4_Tx and 5_TxWk, respectively and 17% of the TSS are annotated as 1_TssA.

**Bottom:** Distribution of annotation for each state in our annotation (left) and in ChromHMM (right). Of the loci annotated as TSS, 72% are known TSS and for ChromHMM 68% of the 1_TssA are known TSS.
4.2 Results

Figure 4.8: RNA-seq read counts in transcription states. RNA-Seq read counts for percentiles for regions classified as transcription. The blue line shows the percentiles for regions classified as Tx Low, Tx Med and Tx High states by our model and the green line shows the read count percentiles for regions classified as Tx and TxWk by ChromHMM model.
Figure 4.9: A cluster-map showing the distribution of the different annotations per cell type.
5 Discussion

Hidden Markov Models were developed since the late 60s, and gained popularity in the 80s when they were extended to support continuous densities.[22] HMMs are also well established and popular tools in the bioinformatics community.[6, 9] In this work we seek to adopt this probabilistic framework for analysis of three different projects with varieties of contexts - from nascent RNA analysis, through analysis of transcription and post transcription mechanisms and ended with analysis of ChIP-seq data for analysis of chromatin annotations. Using data derived from high-throughput experimental procedures, and by translating the biological problems to segmentation and classification problems, we approach them with framework of HMMs: using very simple discrete HMMs, continuous HMMs and multivariate continuous HMMs. With additional analysis of the data and predictions derived from the probabilistic model, we seek to answer some of the research questions.

5.1 Monoubiquitylation of Histone H2B

Our analysis allowed us to estimate the genome-wide elongation rates. Based on this estimation of elongation rates for genes in siRNF20 cells and in control cells (siLacZ), we could not observe a genome wide significant difference in the elongation rates. Even when we compared between group of genes for which there is up-regulation in RNF20 depletion (RNF-suppressed) and a group of down-regulated genes (“RNF20-dependent”), we could not observe a significant difference. This suggest that despite the tight coupling of H2Bub1 to transcription elongation, it is not a rate limiting for transcription elongation in human cells.

As H2Bub1 depletion does not significantly affect the elongation rates, we sought to examine other explanations how H2Bub1 depletion may affect transcription.
H2Bub1 may play a role in Pol II recruitment or regulation of proximal Pol II pausing in gene specific manner. Evidences to such explanation are based on the change in the release time as estimated by the model, where for many genes the release time ($t_0$) was different between siLacZ and siRNF20.

Our analysis provide a novel explanation to the more efficient transcription that appear in RNF20-suppressed genes ([25] and in the mRNA-seq). We suggest that the efficient transcription is due to enhancing the promoter proximal Pol II pausing. This study suggest H2Bub1 role in elongation is yet to be explained well, and further study is required to better understand its function. Other interesting directions for further study may be the role of H2Bub1 in transcription other than coding genes such as transcription associated with enhancers (eRNA).

5.2 Evidences for Stable, Autonomous Uncapped RNA Fragments

Taking advantage of the TEX, CAP-IP and 3 pulldown treatments, we showed their effectiveness to distinguish between different fragments along the transcript, having different read coverage when comparing between the treatments. Using HMM approach we were able to predict and genome-wide annotate cleavage points. Following our further analysis we found evidence for conjecture, that is the downstream cleavage fragment of the 3’UTR remains uncapped and stable as an independent entity from the upstream fragment (coding). Our analysis suggest that it is wide phenomena affecting thousands of genes (more than 2,000 genes show significant coordinated change in read coverage). The phenomena is not limited to specific cell type, and was shown to in U2OS and HEK293 cells.

With more than 30% difference in the read coverage of the tail segment of transcript compared to the body in the different experiments, this phenomena may be important aspect for understanding transcriptomic data: for example miRNA regulation may be less effective if those tails play a role as binding competitors for miRNAs. Additionally, it might by interesting to study cleavage preferences, as well the the stability of the tails, under different conditions (such stress conditions) as they may indirectly influence on the stability of other RNA content in the cells.
The suggest phenomena is yet to be fully characterized. Yet, based on experiments with 3'-end RNA-seq, where we compared distal and the proximal read coverage in transcripts, there are evidences that this process occurs post-transcriptionally, and based on spatial correlation, there are evidences this process may be related to alternative polyadenylation.

Further study of the phenomena may required to fully characterize the mechanism, and explain what are the biochemical processes behind the cleavage. Our results suggest a new category of RNA with a potential of diverse biological properties and call for a new perspective on post-transcriptional regulation of genes via their 3'UTRs.

5.3 Epigenomic Annotator

In this project we suggested a new model for unsupervised annotation of the chromatin using ChIP-seq data. Using this model we derived annotations for more than 100 cell types and showed if can be trained and evaluated efficiently. We showed that the derived model overall agrees with previous studies, by both analysing the derived model parameters themselves - such as the emission probability and by showing functional annotations in the genome are characterized by enrichment of common chromatin patterns.

As the model is multivariate and is free of direct peak calling, it can be sensitive to low signal from various different experiments, which may not be visible if peak calling is taken as a preprocess step. One interesting aspect of this property, is that model being better compared to the popular tool ChromHMM, when we compare the annotation to external data not used for training the model - specifically transcribed states show higher RNA-seq counts. Hence we suggest the model that relays on the raw data, may provide benefit especially in broad peaks common in methylations such as H3K36me3.

The model can be used to derive new annotations and can be directly applied on new ChIP-seq data of other cell types or different cell conditions, for to uncover important genes and regulatory elements which play role in those cell types/conditions. Additonally similar model can be used in future studies and utilized for finding im-
portant loci associated with diseases (GWAS), or combined with other data such as sequence data to derive predictions for enhancers.
Bibliography

[1] NCBI, Epigenomics - Browse Experiments


